Robust structure learning using multivariate T-distributions

Karina Ashurbekova, Sophie Achard, Florence Forbes

Robust structure learning using multivariate t-distributions

Karina Ashurbekova [1,2] & Sophie Achard [2] & Florence Forbes[1]

[1] *Univ. Grenoble Alpes, INRIA, LJK, Mistis team, 38000 Grenoble, France, E-mail firstname.lastname@inria.fr*

[2] *Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France, E-mail firstname.lastname@gipsa-lab.fr*

Résumé. Nous abordons la question de l'apprentissage robuste de la structure de dépendance d'un ensemble de variables continues. Nous considérons l'estimation de matrices de précision creuses qui reflètent une forme de dépendance entre variables. Pour ce faire, nous proposons d'extraire les bonnes caractéristiques de deux méthodes existantes, à savoir *tlasso* et CLIME. La première est basée sur l'observation que la modélisation gaussienne standard aboutit à des procédures qui sont trop sensibles aux valeurs aberrantes et propose l'utilisation de lois de Student comme alternative. Quant à CLIME, il s'agit d'une alternative au populaire Lasso qui peut gérer certaines de ses limitations. Nous combinons ensuite ces idées dans une nouvelle procédure appelée tCLIME qui peut être vue comme une modification de l'algorithme *tlasso*. Les performances de ces procédures sont illustrées sur données simulées et révèlent que tCLIME fonctionne favorablement par rapport aux autres méthodes standard.

Mots-clés. Apprentissage de structure, modèle graphique gaussien, lois de Student, matrice de precision creuse, estimation robuste.

Abstract.

We address the issue of robust graph structure learning in continuous settings. We focus on sparse precision matrix estimation for its tractability and ability to reveal some measure of dependence between variables. For this purpose, we propose to extract good features from existing methods, namely *tlasso* and CLIME procedures. The former is based on the observation that standard Gaussian modelling results in procedures that are too sensitive to outliers and proposes the use of t-distributions as an alternative. The latter is an alternative to the popular Lasso optimization principle which can handle some of its limitations. We then combine these ideas into a new procedure referred to as tCLIME that can be seen as a modified *tlasso* algorithm. Numerical performance is investigated using simulated data and reveals that tCLIME performs favorably compared to the other standard methods.

Keywords. Structure learning, Gaussian graphical model, t-distribution, sparse precision matrix estimation, robust estimation.

1 Introduction

Graphs are an intuitive way of representing and visualizing relationships between variables. In a typical graphical model setting, a p-dimensional random vector $Y = (Y_1, Y_2, \cdots Y_p)$ is represented as an undirected graph denoted by $G = (V, E)$, where V is the set of nodes corresponding to the p variables in Y, and the edge set E describes dependence structure among $Y_1, \cdots Y_p$. In practice, only instances of the variables are observed. The edges encoding the graph structure are unknown and have to be estimated from the observations. Structure or dependence learning has therefore attracted a lot of interest. In a continuous setting, many methods assume the random vector Y is Gaussian with the advantage that for Gaussian vectors, the conditional independence is directly readable from the zeros of the precision matrix. Structure learning can therefore be reduced to standard maximum likelihood parameter estimation. However, in most applications, either the number of observations is too small, or the observations are too noisy, with respect to the dimension of the graph. Most popular approaches face then this issue by resorting to penalized likelihood estimation with the idea of favoring sparse precision matrix estimation. Penalized likelihood approaches based on L_1-norm of the precision matrix Θ include the work of Banerjee et al. [1] and Yuan and Lin [12] while Friedman et al. [6] have developed the most popular method - the *graphical lasso (glasso)*, which is a computationally efficient algorithm that maximizes the penalized log-likelihood function through coordinate descent. As an alternative, Cai et al. [2] have proposed the CLIME procedure, which is a L_1 minimization technique for precision matrix estimation based on different objective function, that tends to provide more sparse solution than *glasso*.

The literature on sparse precision matrix estimation is rapidly growing and received significant attention by the research community. Fan et al. [4], Cai et al. [3] gave a nice review of recent results. Many strong methods are valid only for Gaussian variables such as, for example, SCIO suggested by Sun and Zhang [11] and TIGER of Han and Lie [7]. Although data may deviate from normality in various ways. Outliers and heavy tails frequently occur that can severely degrade the Gaussian models performance. A natural solution is to turn to heavier tailed distributions that remain tractable. In recent papers, Liu et al. [9] exploit the connection between Kendall's tau and Pearson's correlation coefficient in the context of transelliptical distributions to obtain robust estimators of correlation matrices. Zhao and Liu [13] introduced a strong tool for estimation of sparse precision matrix for elliptical family. This method overcome drawbacks of CLIME algorithms. EPIC simultaneously handles data heavy-tailness and allows to calibrate the regularization for estimating different columns of the precision matrix in CLIME algorithm. It uses the combination of the transformed Kendall's tau estimator and Catoni's M-estimator instead of sample covarince matrix. In the special case of multivariate t-distributions Finegold and Drton [5] designed a so-called *tlasso* algorithm to improve graph inference in non Gaussian settings. The *tlasso* algorithm is based on the use of

an Expectation-Maximisation (EM) algorithm that iteratively identifies outlying observations and downweights their impact by using an accordingly modified sample covariance expression. Sparsity is then enforced by replacing the standard M-step by a *glasso* optimization step.

In this work, we propose to modify the approach of Finegold and Drton [5] by replacing the *glasso* M-step with a CLIME M-step, yielding a new procedure for sparse precision matrix estimation referred to as tCLIME. This simple modification enables us to combine benefits from both heavier tail modelling and CLIME optimization. It follows better performance, over *tlasso*, CLIME and EPIC in estimating the zeros of the precision matrix as illustrated on simulations.

This paper is organized as follows. We first describe briefly *glasso* and CLIME optimization problems in Section 1. In Section 5, we recall the *tlasso* algorithm and describe the proposed tCLIME algorithm. Preliminary results on simulations illustrate the superiority of the latter in non Gaussian setting.

2 Gaussian structure learning

For Gaussian graphical models, conditional independence properties are reflected in the precision matrix $\Theta = \{\theta_{jk}\}_{j,k} = \Sigma^{-1}$ of the distribution through zero entries (Lauritzen [8]). Thus inferring the graph in this case corresponds to inferring the nonzero elements of the precision matrix. For each nonedge $(j,k) \notin E$, Y_j and Y_k are conditionally independent given all the remaining $Y_{\setminus\{j,k\}}$ if and only if $\theta_{jk} = 0$.

One of the most commonly used approaches to estimating sparse precision matrix for multivariate normal distribution is through the maximum likelihood. Let $Y_1, \cdots, Y_n$ be observed independent samples from $N(\mu, \Sigma)$, $\Theta = \Sigma^{-1} \in \mathbb{R}^{p\times p}$ is the precision matrix and $S = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)(Y_i - \mu)^T$ is the sample covariance matrix. The negative Gaussian log-likelihood is given by: $l(\Theta) = tr(S\Theta) - log\ det(\Theta)$. In penalized likelihood methods a L_1-norm penalty is added to the log-likelihood function to favor zero entries in Θ. One of the most popular techniques to maximize the resulting L_1-penalized log-likelihood function is *glasso* (Friedman et al. [6]):

$$\min_{\Theta \succ 0} L(\Theta) = tr(\Theta S) - \log\ det(\Theta) + \rho\|\Theta\|_1 \tag{1}$$

where ρ is a regularization parameter. Larger values of ρ correspond to larger levels of sparsity in Θ. If ρ is equal to zero then the solution of the (1) is the inverse of the sample covariance matrix.

An alternative to *glasso* is a non-likelihood based approach called CLIME which solves instead:

$$\begin{aligned} &\min\ \|\Theta\|_1 \\ &\text{s.t.: } |S\Theta - I|_\infty \leqslant \rho, \end{aligned} \tag{2}$$

where ρ is a tuning parameter and S is the sample covariance matrix. No symmetry condition is imposed on Θ and symmetry is achieved by setting both estimates $\widehat{\theta}_{ij}$ and $\widehat{\theta}_{ji}$ to the one with the smaller magnitude. Note that theoretical justification of CLIME algorithm heavily relies on the Sub-Gaussian tail assumption. However CLIME, as *glasso*, still uses the usual sample covariance matrix as input which makes it sensitive to outliers.

As explained in Cai et al. [2], the CLIME motivation and its relationship to *glasso* comes from the following observation. The solution of the *glasso* optimization problem:

$$\widehat{\Theta}_{glasso} = \min_{\Theta \succ 0} \; tr(\Theta S) - \log \; det(\Theta) + \rho \|\Theta\|_1 \tag{3}$$

satisfies $(\widehat{\Theta}_{glasso})^{-1} - S = \rho \widehat{Z}$, where $\widehat{Z}$ is an element of the subdifferential $\partial \|\widehat{\Theta}_{glasso}\|_1$. This leads to the optimization problem:

$$\begin{aligned} &\min \; \|\Theta\|_1 \\ &\text{s.t.: } |\Theta^{-1} - S|_\infty \leqslant \rho, \end{aligned} \tag{4}$$

After multiplying the constraint with Θ we find the problem (2).

3 T-distribution graphical models: *tlasso* and tCLIME

Among the various existing forms of the multivariate t-distribution (Nadarajah and Kotz [10]), the most common form with parameters μ, Σ and ν is given by:

$$t_p(y; \mu, \Sigma, \nu) = \frac{\Gamma(\frac{\nu+p}{2})}{(\pi\nu)^{p/2}\Gamma(\frac{\nu}{2})|\Sigma|^{1/2}} \big[1 + \frac{\delta(y, \mu, \Sigma)}{\nu}\big]^{-(\nu+p)/2}$$

where $\delta(y, \mu, \Sigma) = (y-\mu)^T\Sigma^{-1}(y-\mu)$ is the Mahalanobis distance between y and μ, $y \in \mathbb{R}^p$. The vector $\mu \in \mathbb{R}^p$ and the positive definite matrix Σ determine the first two moments of the distribution when $\nu > 2$. The covariance matrix of the t-distribution is $\nu/(\nu-2)\Sigma$. We will keep the convenient notation of Finegold and Drton [5] and denote by $\Theta = \Sigma^{-1}$ to draw a parallel between Gaussian graphical models and graphical models based on t-distribution. Unlike the Gaussian case, in a t-distribution $\theta_{jk} = 0, j \neq k$ no longer corresponds to conditional independence of Y_j and Y_k given all the remaining $Y_{\setminus\{j,k\}}$. However, despite the lack of conditional independence, the conditional uncorrelation property still holds.

Mimicking the *glasso* idea, Finegold and Drton [5] proposed to add a L1-norm penalty on Θ to the maximization of the log-likelihood function in the t-distribution case. It follows a *tlasso* procedure that is more robust to outliers. Inference is made using an EM algorithm: τ is a hidden variable and conditional distribution of Y given τ is $N_p(\mu, \Sigma/\tau)$. Degree of freedom parameter ν is assumed to be known. At the $(t+1)$th M-step, an update value $\Theta^{(t+1)}$ is found by solving the following optimization problem:

$$\min_{\Theta \succ 0} \; tr(\Theta S_{\tau^{(t+1)}YY}(\mu^{(t+1)})) - \log|\Theta| + \rho\|\Theta\|_1, \tag{5}$$

where $\tau_i^{(t+1)}$ is the conditional expectation of τ_i calculated in the $(t+1)$st E-step as $\tau_i^{(t+1)} = \frac{\nu+p}{\nu+\delta_Y(\mu^{(t)},\Sigma^{(t)})}$. A "weighted sample covariance matrix" is then computed as:

$$S_{\tau^{(t+1)}YY}(\mu) = \frac{1}{n}\sum_{i=1}^{n} \tau_i^{(t+1)}(Y_i - \mu^{(t+1)})((Y_i - \mu^{(t+1)}))^T \tag{6}$$

and $\mu^{(t+1)} = \sum_{i=1}^n(\tau_i^{(t+1)}Y_i)/\sum_{i=1}^n \tau_i^{(t+1)}$. Note that (5) is a similar objective function minimized by *glasso*, so that the proposed *tlasso* reduces to solving at each iteration a *glasso* optimization.

In the same EM framework, we propose to replace *glasso* by CLIME for its better performance in the M-step. It follows that (5) becomes:

$$\begin{aligned} &\min \ \|\Theta\|_1 \\ &\text{s.t.: } |S_{\tau^{(t+1)}YY}(\mu^{(t+1)})\Theta - I|_\infty \leqslant \rho, \end{aligned} \tag{7}$$

We refer to the resulting procedure as tCLIME. As *tlasso*, tCLIME uses the "weighted sample covariance matrix" instead of the sample covariance matrix, which makes both methods less sensitive to outliers in comparison to their Gaussian counterparts (*glasso* and CLIME).

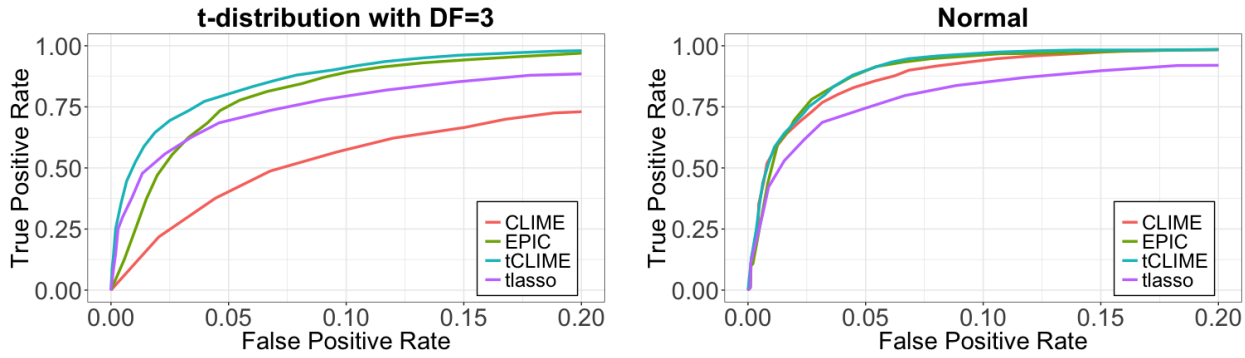

Figure 1: ROC curves illustrating the performance of tlasso, CLIME, tCLIME and EPIC methods on 2 different data sets. The tuning parameter ρ is chosen in the range $[0.1, 2.5]$ with stepsize 0.05 for *tlasso* and $[0.01, 0.4]$ with stepsize 0.01 for CLIME, tCLIME anf EPIC.

4 Simulation results

Since tCLIME is a combination of *tlasso* and CLIME, the tCLIME performance is compared to that of *tlasso* and CLIME. We also compare tCLIME with the EPIC method which is a strong estimator of sparse precision matrix for elliptical distributions. We

generate a random 100×100 sparse precision matrix Θ according to the procedure described in Finegold and Drton [5] and simulate n=150 observations from $t_{100}(0, \Theta^{-1}, 3)$ and $N_{100}(0, \Theta^{-1})$. The four algorithms are then run with different values of ρ. To measure how well the sparsity of the true precision matrix is recovered, the whole process is repeated 50 times. The corresponding ROC curves are shown in Figure 1. For Gaussian data, the tCLIME and EPIC performance is similar and better than that of *tlasso* and CLIME. When data are generated from a t-distribution, tCLIME significantly outperforms CLIME and also shows better results than EPIC and *tlasso*.

5 Conclusion

We have introduced tCLIME a modified version of the *tlasso* algorithm. tCLIME provides competitive results and can be used for robust sparse precision matrix estimation. However further simulations should be made to account for higher dimensional cases, as well as tests on real data.

Bibliographie

[1] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res.*, 9(Mar):485–516, 2008.

[2] T. Cai, W. Liu, and X. Luo. A constrained l-1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.*, 106(494):594–607, 2011.

[3] T.T. Cai, Z. Ren, and H.H. Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Stat.*, 10(1):1–59, 2016.

[4] J. Fan, Y. Liao, and H. Liu. An overview of the estimation of large covariance and precision matrices. *Econom. J.*, 19 (1), 2016.

[5] M. Finegold and M. Drton. Robust graphical modeling of gene networks using classical and alternative t-distributions. *Ann. Appl. Stat.*, pages 1057–1080, 2011.

[6] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[7] L. Han and W. Lie. Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. *arXiv preprint arXiv:1209.2437*, 2012.

[8] S.L. Lauritzen. *Graphical Models*. Oxford Univ. Press, New York, 1996.

[9] H. Liu, F. Han, and C. Zhang. Transelliptical graphical models. In *Adv. Neural Inf. Process Syst.*, pages 800–808, 2012.

[10] S. Nadarajah and S. Kotz. *Multivariate t Distributions and their Applications*. Cambridge, 2004.

[11] T. Sun and C.H. Zhang. Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.*, 14(1):3385–3418, 2013.

[12] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[13] T. Zhao and H. Liu. Calibrated precision matrix estimation for high-dimensional elliptical distributions. *IEEE Trans. Inf. Theory*, 60(12):7874–7887, 2014.